Bandits: An Introduction

INF581 Advanced Machine Learning and Autonomous Agents

Jesse Read

Last Updated: January 19, 2022

Outline

Introduction to Bandits

Bandits

We consider stochastic multi-arm bandits.

Motivating Example: Adaptive Drug Trials

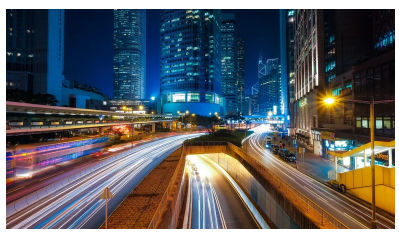
A doctor, specialising in treatment of a particular disease, can prescribe one of K possible drugs (bandit arms). S/he sees a patient, assigns a drug (takes an action/pulls a bandit arm), and observes the effectiveness (reward), then proceeds to the next patient, and thus in **sequence**. The goal is to cure as many patients as possible.

For patient $t = 1, 2, \ldots, T$:

1. Give drug $A_t \in \{1, \ldots, K\}$ (action) to patient t
2. Observe outcome (reward) $R_t \in \{\text{recovered}, \neg\text{recovered}\}$
3. Update knowledge about the estimated outcome

Bandit Applications

- Medical trials (prescribe drugs; maximize health outcome)
- Finance (invest in stock; maximize payoff)
- Influence Maximization (select influencer; maximize influence)
- Services (select provider to use; maximize service quality)
- Recommender systems (which movie to watch; acceptance of recommendation)
- Marketing (which ad. to display; maximize revenue from ads)
- Robotics and logistics (select strategy; maximize productivity)

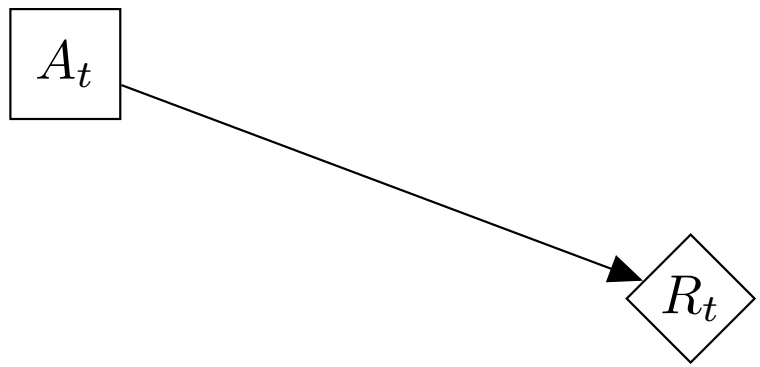





Bandits as an Influence Diagram

An Influence Diagram is an extension to a Bayesian Network, with decision options (square nodes), and rewards/value they incur (diamond nodes).

We take an action $A_t \in \{1, \ldots, K\}$ and receive reward $R_t \in \mathbb{R}$:



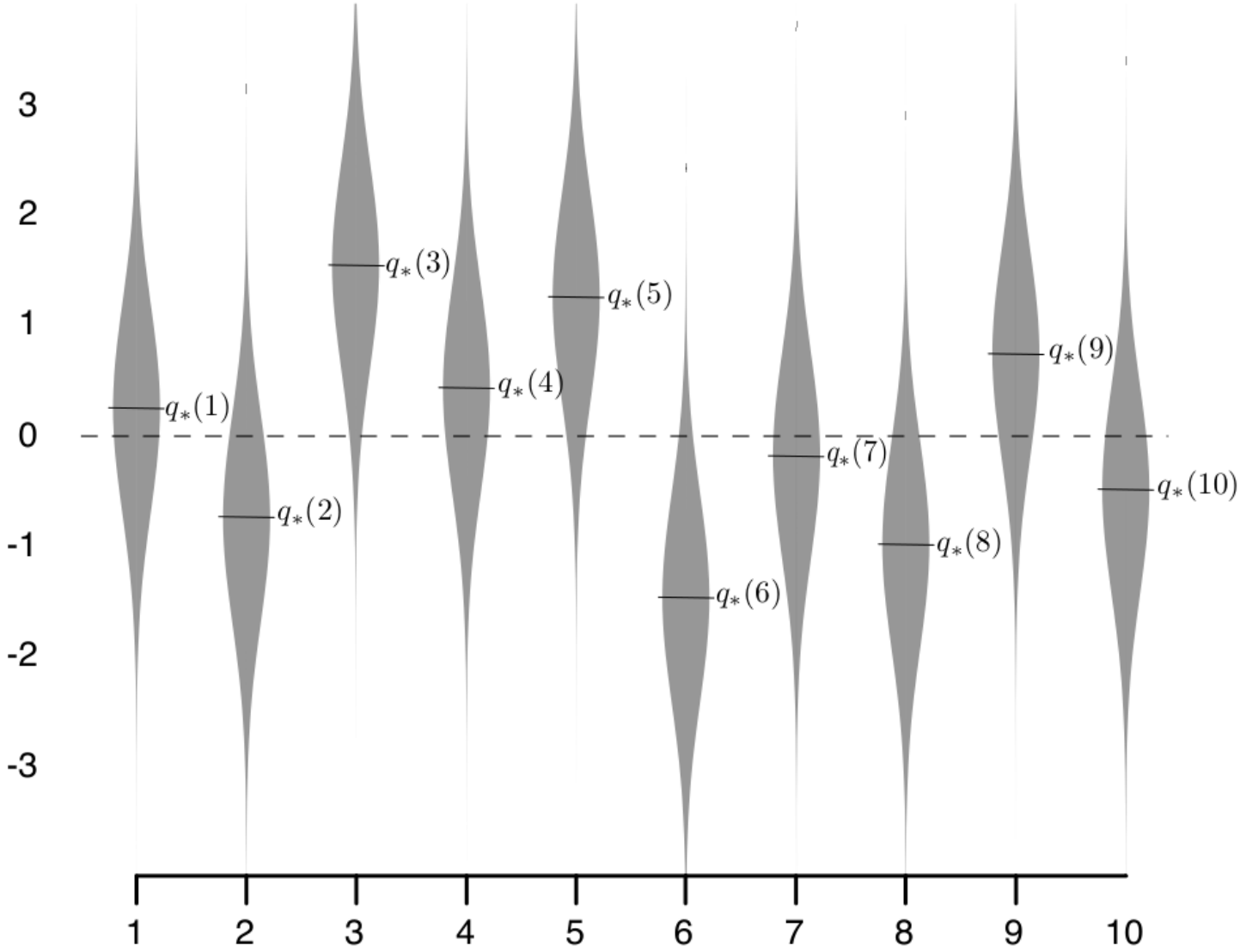
$$\nu(R_t|A_t)\pi(A_t)$$

Let $\nu(R_t|A_t = a) \equiv \nu_a(R_t)$. We want to learn about distributions $\nu_1, \ldots, \nu_K$ and, in particular, each

$$q_*(a) = \mathbb{E}_{R_t \sim \nu_a}[R_t|A_t = a]$$

(the expected value of pulling arm a).

An example ($K = 10$; where $q_*(a) = \mathbb{E}[R_t | A_t = a]$):



Sutton and Barto. *Reinforcement Learning: An Introduction*, 2020.

Sequential Decision Making

Earlier: For each $t = 1, \ldots, T$ (in test set) the 'agent'/model f:

1. Observes X_t
2. Takes 'action' $\widehat{Y}_{t,1}, \widehat{Y}_{t,2}, \ldots, \widehat{Y}_{t,L} = f(X)$
3. [Maybe] receives Y_t; Update our model f with (X_t, Y_t)

Generic bandits framework: For $t = 1, \ldots, T$, the agent π

1. Takes action $A_t \sim \pi$
2. Observes reward R_t (Note: this R_t *only* wrt A_t)
3. Update our model ν, agent π, with (A_t, R_t)

Sequential Decision Making

Earlier: For each $t = 1, \ldots, T$ (in test set) the 'agent'/model f:

1. Observes X_t
2. Takes 'action' $\widehat{Y}_{t,1}, \widehat{Y}_{t,2}, \ldots, \widehat{Y}_{t,L} = f(X)$
3. [Maybe] receives Y_t; Update our model f with (X_t, Y_t)

Generic bandits framework: For $t = 1, \ldots, T$, the agent π

1. Takes action $A_t \sim \pi$
2. Observes reward R_t (Note: this R_t *only* wrt A_t)
3. Update our model ν, agent π, with (A_t, R_t)

Important difference: Above, we cannot affect observation X_{t+1} at time t, in any way with any decision. Below, our goal at time t is not [only] to optimize R_t, rather to optimize: $\mathbb{E}[R_t, R_{t+1}, \ldots, R_T]$. This is not obtained with a single decision but a sequence of decisions.

Reward and Regret

The mean reward (expected reward) of an arm at time t is

$$\mu(a) := \mathbb{E}_{R_t \sim \nu_a}[R_t | A_t = a]$$

The best result (via highest values/rewards) is obtained with

$$\mu^* := \max_{a \in \mathcal{A}} \mu(a) \qquad a^* = \operatorname*{argmax}_{a \in \mathcal{A}} \mu(a)$$

(the largest mean of all the arms).

At round T, the regret for our policy is

$$\mathcal{R}_T = \mu^* T - \mathbb{E}_\pi \left[\sum_{t=1}^{T} R_t \right]$$

We want to minimize this. Intuition: play the bandit such that in hindsight, we couldn't have done any better.

Remark: rv $\mathcal{R}_T(\pi, \nu)$ inherits randomness from ν and π.

Exploitation vs Exploration

- Exploit: Take greedy actions (best under current knowledge).
- Explore: Improve knowledge for the future (i.e., learning).

The dilemma: We cannot do both always. It's a trade-off.

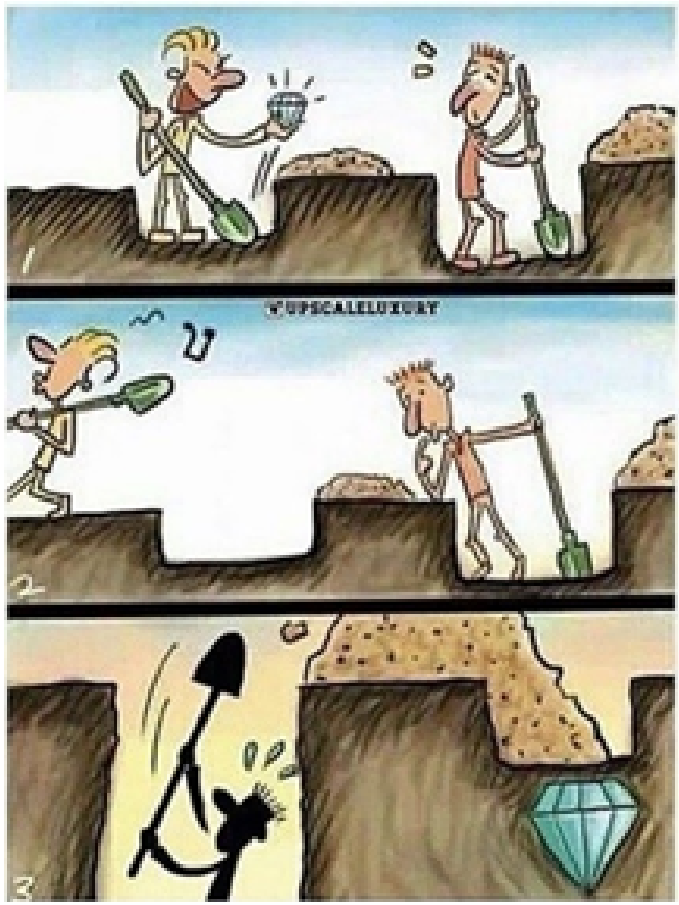


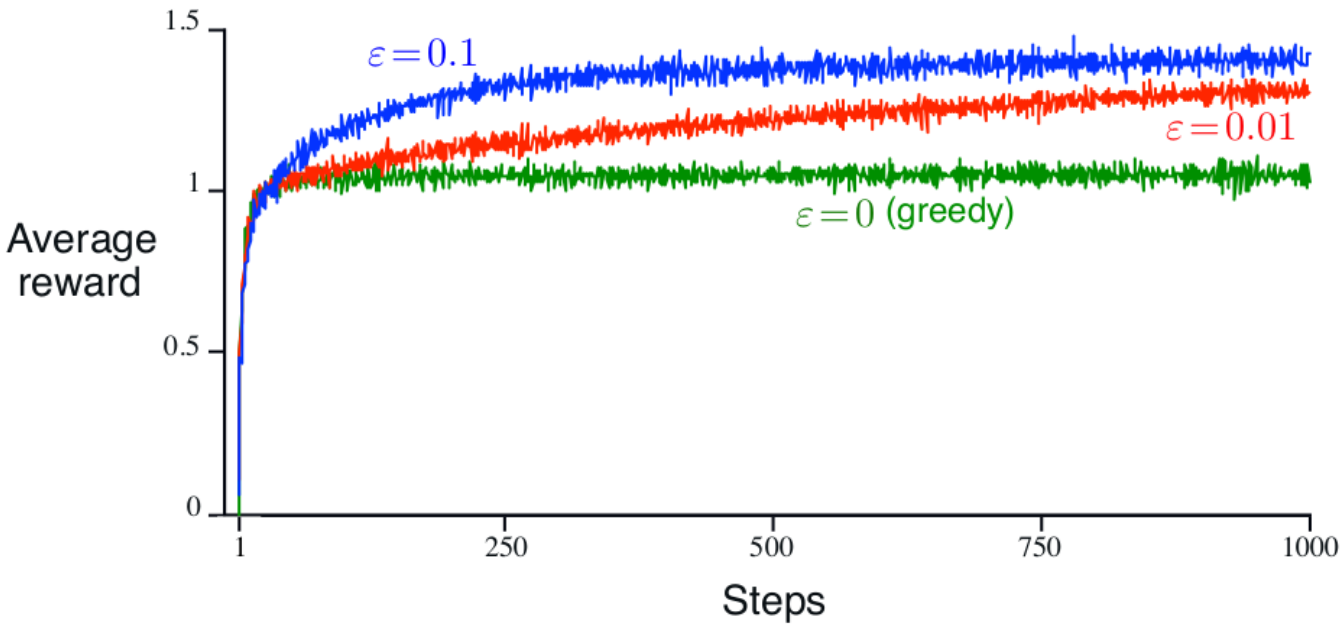
Image Source: [1]

The ϵ-Greedy Algorithm

ϵ-Greedy[1]: play the best arm with probability $1 - \epsilon$ (greedy action):

$$\hat{a}^* = \underset{a \in \mathcal{A}}{\text{argmax}}\, \hat{\mu}(a)$$

and (else) play a random arm $a \sim \mathcal{A}$ with probability ϵ (explore). Then, we can update $\hat{\mu}(\cdot)$ for the arm played.



ϵ-greedy vs greedy. Ref: Sutton and Barto. *Reinforcement Learning: An Introduction*, 2020.

[1] N.B. Related to ϵ-approximate tree search, but it's not the same!

The Upper Confidence Bounds (UCB) Algorithm

Problems with ϵ-Greedy

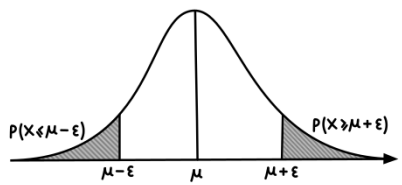
- ϵ may be too large (over-exploring) or too small (over-greedy)
- algorithm may be biased by initial estimates of $\hat{\mu}(a)$

Tail Probabilities

Suppose iid samples from arm a: $R_1, R_2, \ldots, R_T$.
Our unbiased estimate $\hat{\mu}$ (empirical mean over T draws):

$$\hat{\mu} = \frac{1}{T} \sum_{t=1}^{T} R_t$$

How confident can we be that the true mean μ is not greater?



(we're interested in the tail probabilities). We can derive[2]:

$$P\left(\mu \geq \hat{\mu} + \sqrt{\frac{2\log(1/\delta)}{T}}\right) \leq \delta$$

[2] See, e.g., Lattimore and Szepesvári, Bandit Algorithms, 2020. Chap. 5.

Upper-Confidence-Bounds (UCB)

Recall: Only one arm can be pulled per round!
Let $N_k(t)$ be the the number of samples from the k-th arm seen by time-step t, with an empirical mean of $\hat{\mu}_k(t)$.

We can define, for time-step t:

$$\text{UCB}_k(t-1,\delta) = \hat{\mu}_k(t-1) + \sqrt{\frac{2\log(1/\delta)}{N_k(t-1)}}$$

where $\text{UCB}_k(t-1,\delta) = \infty$ when $N_k(t-1) = 0$.

- Optimism in the face of uncertainty: assume largest plausible value when we haven't chosen the arm yet.
- Essentially two terms: exploitation + exploration
- Optimism will decay with experience.

with confidence level δ.

The UCB Algorithm

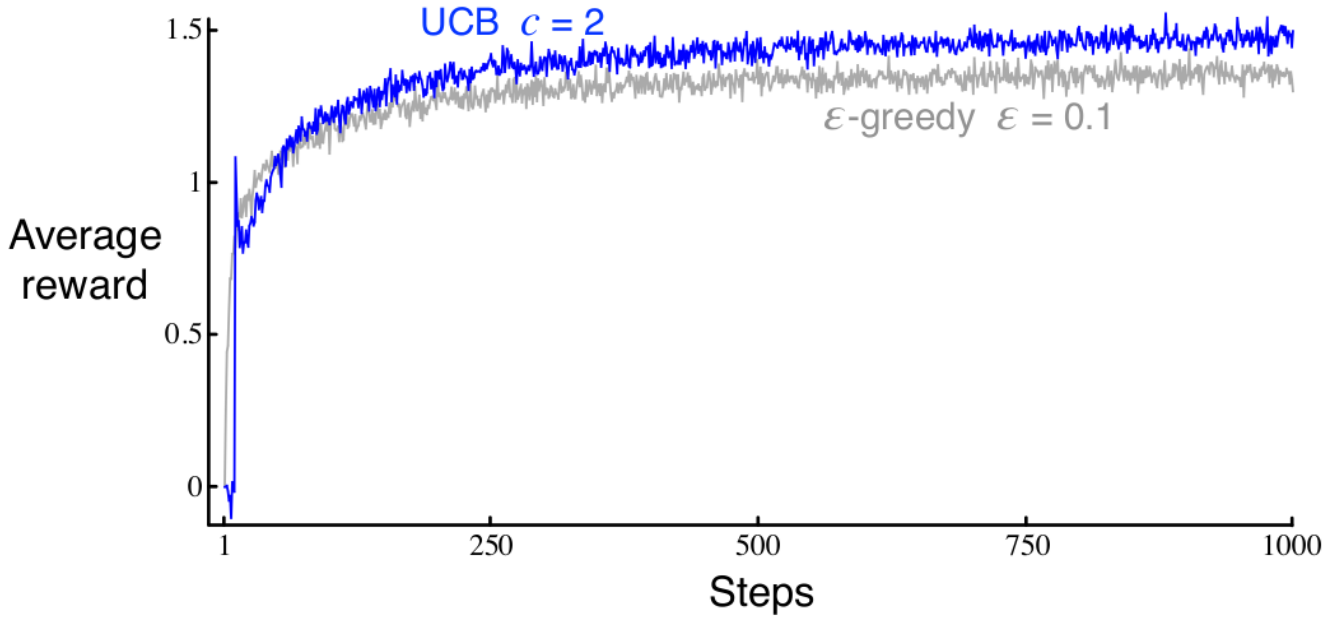
1: **procedure** UCB ALGORITHM(δ)
2: Init. $\{N_k(0)\}_{k=1}^K = 0$ and $\{\text{UCB}_k(0,\delta)\}_{k=1}^K = \infty$
3: **for** $t = 1, \ldots, T$ **do**
4: Choose action

$$A_t = \underset{k \in \mathcal{A}}{\text{argmax}}\, \text{UCB}_k(t-1, \delta)$$

5: Pull the arm A_t and observe reward R_t
6: Update UCB_k: $N_{A_t}(t)$ and $\hat{\mu}_k(t)$

for our choice of $\delta \in (0,1)$: the confidence level which quantifies the degree of uncertainty. It should be small enough to ensure reasonable optimism, but not so large as to encourage too much exploration of suboptimal arms.

General idea: Choosing the arm with the largest UCB implies only playing only arms k where the true mean μ_k could reasonably be larger than the arms which have been played often.



Sutton and Barto. *Reinforcement Learning: An Introduction*, 2020.

Summary (So Far)

Bandits . . .

. . . imply sequential decision making: a decision we make now affects the observation we get next, and so on, affecting our loss (regret) potentially far from now.

. . . could be also be called ‘reinforcement learning from the same state’ or ’. . . without a change in state’; we only observe a reward.

. . . have many real-world applications.

. . . involve the all-important Exploration vs Exploitation tradeoff.

Bandits: An Introduction

INF581 Advanced Machine Learning and Autonomous Agents

Jesse Read